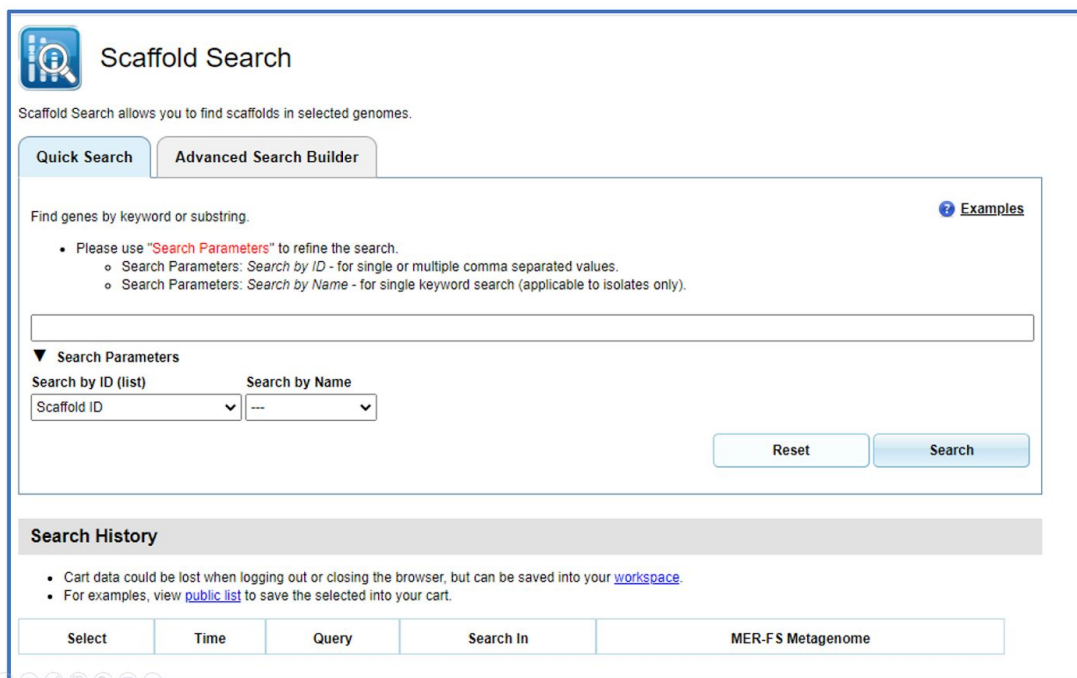


Scaffold Search User Guide

(9/28/2020)

Introduction

IMG has developed a new **Scaffold Search** feature under **Find Genomes** (Figure 1). Similar to other new search functions in IMG, the new Scaffold Search feature allows IMG users to perform quick search using IDs, or to search IMG scaffolds using a more advanced query builder.



Scaffold Search

Scaffold Search allows you to find scaffolds in selected genomes.

Quick Search | **Advanced Search Builder**

Find genes by keyword or substring. [Examples](#)

- Please use "Search Parameters" to refine the search.
 - Search Parameters: Search by ID - for single or multiple comma separated values.
 - Search Parameters: Search by Name - for single keyword search (applicable to isolates only).

▼ Search Parameters

Search by ID (list) Search by Name

Scaffold ID ---

Reset Search

Search History

- Cart data could be lost when logging out or closing the browser, but can be saved into your [workspace](#).
- For examples, view [public list](#) to save the selected into your cart.

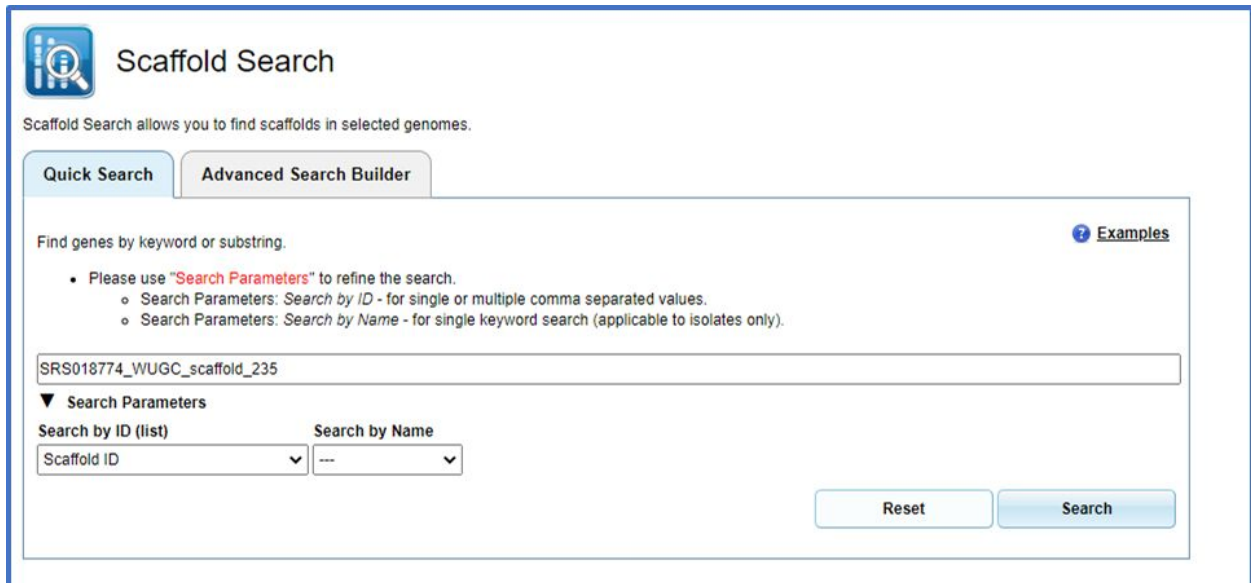
Select	Time	Query	Search In	MER-FS Metagenome
--------	------	-------	-----------	-------------------

Quick Search

Quick search allows users to search scaffolds using IMG scaffold IDs, external accessions or scaffold names. Note that metagenome scaffolds can only be searched using IMG IDs.

Example 1:

Suppose a user is interested in finding a scaffold with ID “SRS018774_WUGC_scaffold_235” from HMP study, but does not know the corresponding metagenome OID. The user can simply use the Scaffold ID search option in Quick Search:



Scaffold Search

Scaffold Search allows you to find scaffolds in selected genomes.

Quick Search | **Advanced Search Builder**

Find genes by keyword or substring. [? Examples](#)

- Please use “**Search Parameters**” to refine the search.
 - Search Parameters: *Search by ID* - for single or multiple comma separated values.
 - Search Parameters: *Search by Name* - for single keyword search (applicable to isolates only).

SRS018774_WUGC_scaffold_235

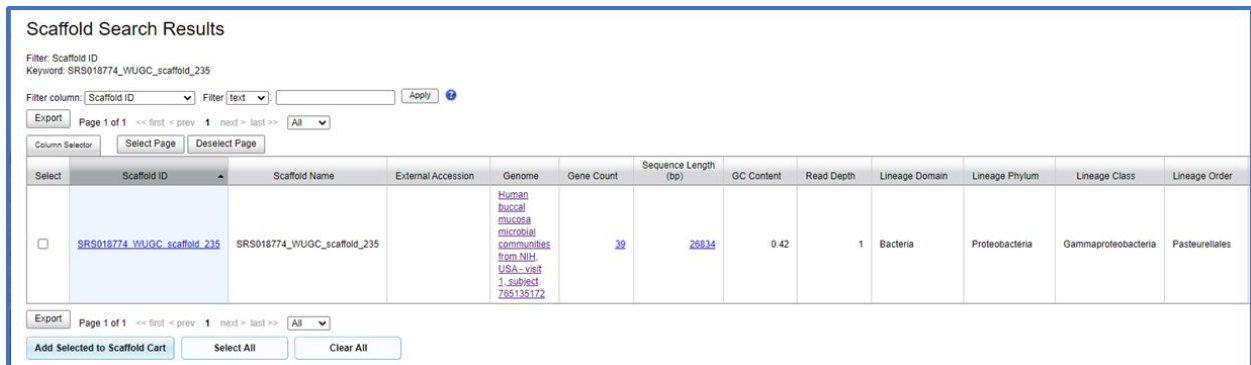
▼ **Search Parameters**

Search by ID (list) Search by Name

Scaffold ID ---

Reset **Search**

Enter the scaffold ID in the blank, and click the **Search** button, the following result will appear:



Scaffold Search Results

Filter: Scaffold ID
Keyword: SRS018774_WUGC_scaffold_235

Filter column: Scaffold ID Filter (text) Apply

Export Page 1 of 1 << first < prev 1 next > last >> All

Column Selector Select Page Deselect Page

Select	Scaffold ID	Scaffold Name	External Accession	Genome	Gene Count	Sequence Length (bp)	GC Content	Read Depth	Lineage Domain	Lineage Phylum	Lineage Class	Lineage Order
<input type="checkbox"/>	SRS018774_WUGC_scaffold_235	SRS018774_WUGC_scaffold_235		Human buccal microbial communities from NIA USA - visit 765135172	39	26834	0.42	1	Bacteria	Proteobacteria	Gamma proteobacteria	Pasteureliales

Export Page 1 of 1 << first < prev 1 next > last >> All

Add Selected to Scaffold Cart Select All Clear All

The user can select the scaffold and add it to Scaffold Cart for further analysis.

Advanced Search

Advanced Search Builder allows IMG users to form a more advanced search using a combination of any of the following fields:

- Scaffold Taxonomy
 - Domain
 - Phylum
 - Class

- Order
- Family
- Genus
- Species
- Function IDs
 - COD ID (list)
 - KOG ID (list)
 - Pfam ID (list)
 - TIGRfam ID (list)
 - SMART ID (list)
 - SUPERFam ID (list)
 - CATH FunFam ID (list)
 - KO ID (list)
 - Enzyme ID (list)
 - IMG Term ID (list)
- Function Names
 - COD Name
 - KOG Name
 - Pfam Name
 - TIGRfam Name
 - SMART Name
 - SUPERFam Name
 - CATH FunFam Name
 - KO Name
 - Enzyme Name
 - IMG Name
- Scaffold Statistics
 - Scaffold Topology
 - Scaffold Nucleotide Length
 - Scaffold GC Percentage
 - Scaffold Read Depth
 - Scaffold Gene Count
 - Scaffold CDS Gene Count
 - Scaffold tRNA Gene Count
 - Scaffold 5S rRNA Gene Count
 - Scaffold 16S rRNA Gene Count
 - Scaffold 18S rRNA Gene Count
 - Scaffold 23S rRNA Gene Count
 - Scaffold 28S rRNA Gene Count
 - Scaffold other rRNA Gene Count

Example 2:

Suppose a user has two marine microbial communities from Delaware Coast in the Genome Cart (IMG taxon IDs : 3300000101, 3300000116):

Genome Cart

Genomes in Cart | Upload & Export & Save

[Group Genome Cart by Phyla](#)

hint: Scaffolds will not be added into cart for very large genomes. Only scaffolds (assembled data only) of selected MER-FS genomes can be added into cart.

Add Scaffolds of Selected Genomes to Cart | Toggle Selected | Select All | Clear All | Remove Selected

Filter column: Domain | Filter text: | Apply

Export | Page 1 of 1 | << first < prev 1 next > last >> | All

Column Selector | Select Page | Deselect Page

Select	Domain	Sequencing Status	Study Name	Genome Name / Sample Name	Sequencing Center	IMG Genome ID	Is Public	Genome Size * assembled	Gene Count * assembled
<input type="checkbox"/>	*Microbiome	Draft	Marine microbial communities from Delaware Coast	Marine microbial communities from Delaware Coast sample from Delaware MO Early Summer May 2010	DOE Joint Genome Institute (JGI)	3300000101	Yes	647909234	1375242
<input type="checkbox"/>	*Microbiome	Draft	Marine microbial communities from Delaware Coast	Marine microbial communities from Delaware Coast sample from Delaware MO Spring March 2010	DOE Joint Genome Institute (JGI)	3300000116	Yes	590073671	1417215

Export | Page 1 of 1 | << first < prev 1 next > last >> | All

To find all scaffolds with both 16s and 23s rRNAs in these two metagenomes, the user can use the Advanced Search Builder with the following two conditions:

- Scaffold Statistics -- Scaffold 16S rRNA Gene Count * (Range) >= 1
- Scaffold Statistics -- Scaffold 23S rRNA Gene Count * (Range) >= 1

Then add both metagenomes in the cart to search.

Quick Search

Advanced Search Builder

hint:

If timed out while running, please take the following measures: 1) Make choice of 'Selected Genomes' instead of 'All Isolates'. 2) Reduce the number of selection for 'Selected Genomes'.

Find Scaffs by constructing a query using keywords, substrings, and AND/(AND NOT)/OR operators.

- The 'AND' / 'AND NOT' query operator(s) that combine builder lines will be processed sequentially.
- Range queries enabled with keyword 'to' indicate range searches.
For example, input '1 to 1000' in the field of Scaffold Statistics-> Scaffold Nucleotide Length will retrieve datasets collected at length between 1 and 1000.
- Mathematical operators '>=', '>', '<', '<=' can be applied to a range search.
For example, input '<=1000' in the field of Scaffold Nucleotide Length will retrieve datasets collected at length of less than 1000.
- Use builder line Scaffold Statistics-> Scaffold 16S rRNA Gene Count to retrieve 16S Sequences, or similarly other rRNA sequences of interest.
- The maximum number of builder lines is limited to 5.
- Click on "Evaluate Query" button to see the results of each specific subquery and the overall count. Click on "Search" button to see the overall results.

Scaffold Statistics

Scaffold 16S rRNA Gene Count * (Range)

>= 1

-

Remove

AND

Scaffold Statistics

Scaffold 23S rRNA Gene Count * (Range)

>= 1

-

Remove

+

Add new builder line

Search in:

☒ Selected Genomes
 ☐ All Isolates

- The total selection for 'Selected Genomes' can only be less than 50 metagenomes.
- There could be a significant delay if more than 1000 genomes selected or if 'All Isolates' selected.

Sequencing Status

Domain

All Finished, Permanent Draft and Draft

Genome Cart

☒ List
 ☐ Tree

Show

Search for:

<enter a genome name to search>

Marine microbial communities from Delaware Coast, sample from Dela

Marine microbial communities from Delaware Coast, sample from Dela

Add >

Add All >>

< Remove

<< Remove All

Selected Genomes

2 selected

Marine microbial communities from Delaware Coast,

Marine microbial communities from Delaware Coast,

MER-FS Metagenome:

Assembled

Constructed Query: Use the builder above to create search query.

Reset

Evaluate Query

Search

Click the **Search** button, and the following result with 91 scaffolds will be displayed:

Advanced Scaffold Search Results

MER-FS Metagenome: assembled

Query: (Scaffold Statistics -- Scaffold 16S rRNA Gene Count * (Range) [>= 1]) AND (Scaffold Statistics -- Scaffold 23S rRNA Gene Count * (Range) [>= 1])

Selected Genomes: 2

- (Scaffold Statistics -- Scaffold 16S rRNA Gene Count * (Range) [>= 1]: 561 count(s).
- (Scaffold Statistics -- Scaffold 23S rRNA Gene Count * (Range) [>= 1]: 954 count(s). (Earlier results channelled into this search.)

Final Combination: 91 count(s)

Filter column: Scaffold ID Filter: text Apply

Export Page 1 of 1 << first < prev 1 next > last >> All

Select	Scaffold ID	Scaffold Name	External Accession	Genome	Gene Count	Sequence Length (bp)	GC Content	Read Depth	Lineage Domain	Lineage Phylum	Lineage Class	Lineage Order
<input type="checkbox"/>	DeiMOSpr2010_c10000493	DeiMOSpr2010_c10000493		Marine microbial communities from Delaware Coast sample from Delaware MO Spring March 2010	24	25292	0.51	113	Bacteria	Verrucomicrobia	Opilutae	Puniceococcales
<input type="checkbox"/>	DeiMOSpr2010_c10000558	DeiMOSpr2010_c10000558		Marine microbial communities from Delaware Coast sample from Delaware MO Spring March 2010	37	23755	0.26	175				
<input type="checkbox"/>	DeiMOSpr2010_c10000855	DeiMOSpr2010_c10000855		Marine microbial communities from Delaware Coast sample from Delaware MO Spring March 2010	5	5327	0.50	183				

Again, scaffolds in the result table can be selected to add to the Scaffold Cart for further analysis. The user can also add any additional attributes in the Table Configuration and click the **Redisplay** button to add the additional fields to the table display.

Example 3:

Suppose the user is interested in finding all scaffolds in the same two metagenomes satisfying the following 3 conditions:

- Having genes annotated with either pfam00543 (P-II - Nitrogen regulatory protein P-II) or pfam00909 (Ammonium_transp - Ammonium Transporter Family)
- Scaffold lineage assigned to the *Proteobacteria* phylum
- With length greater than or equal to 10,000 bp

The user can use the Advanced Search Builder with the following 3 builder lines:

- Function IDs -- Pfam ID (list) *: pfam00543, pfam00909
- Scaffold Taxonomy -- Phylum: Proteobacteria
- Scaffold Statistics -- Scaffold Nucleotide Length * (Range): >= 10000

And add the 2 metagenomes to search.

Function IDs

Pfam ID (list) *

pfam00543, pfam00909

- Remove

AND

Scaffold Taxonomy

Phylum

Proteobacteria

- Remove

AND

Scaffold Statistics

Scaffold Nucleotide Length * (Range)

>= 10000

- Remove

+ Add new builder line

Search in:

☒ Selected Genomes
 ☐ All Isolates

- The total selection for 'Selected Genomes' can only be less than 50 metagenomes.
- There could be a significant delay if more than 1000 genomes selected or if 'All Isolates' selected.

Sequencing Status

Domain

All Finished, Permanent Draft and Draft

Genome Cart

List

Tree

Show

Search for:

<enter a genome name to search>

Marine microbial communities from Delaware Coast, sample from Delav

Marine microbial communities from Delaware Coast, sample from Delav

Add >

Add All >>

< Remove

<< Remove All

3300000101

3300000116

2 selected

MER-FS Metagenome: Assembled

Constructed Query:

(Function IDs -- Pfam ID (list) * [pfam00543, pfam00909]) AND (Scaffold Taxonomy -- Phylum [Proteobacteria]) AND (Scaffold Statistics -- Scaffold Nucleotide Length * (Range) [>= 10000])

Reset

Evaluate Query

Search

Hide All Categories

The result shows that 68 scaffolds satisfying the condition:

Advanced Scaffold Search Results										
MER-FS Metagenome: assembled										
Query: (Function IDs -- Pfam ID (list) * [pfam00543, pfam00909]) AND (Scaffold Taxonomy -- Phylum [Proteobacteria]) AND (Scaffold Statistics -- Scaffold Nucleotide Length * (Range) [>= 10000])										
Selected Genomes: 2										
<ul style="list-style-type: none"> (Function IDs -- Pfam ID (list) * [pfam00543, pfam00909]): 1092 count(s). (Scaffold Taxonomy -- Phylum [Proteobacteria]): 360555 count(s). (Earlier results channelled into this search.) (Scaffold Statistics -- Scaffold Nucleotide Length * (Range) [>= 10000]): 7375 count(s). (Earlier results channelled into this search.) 										
Final Combination: 68 count(s).										
<div>Add Selected to Scaffold Cart</div> <div>Select All</div> <div>Clear All</div>										
Filter column: Scaffold ID Filter text Apply										
Export Page 1 of 1 << first < prev 1 next > last >> All										
Column Selector Select Page Deselect Page										
Select	Scaffold ID	Scaffold Name	External Accession	Genome	Gene Count	Sequence Length (bp)	GC Content	Read Depth	Lineage Domain	Lineage Phylum
<input type="checkbox"/>	DeMOSpr2010_c10000002	DeMOSpr2010_c100000002		Marine microbial communities from Delaware Coast, sample from Delaware MO Spring March 2010	137	148253	0.35	207	Bacteria	Proteobacteria
<input type="checkbox"/>	DeMOSpr2010_c10000059	DeMOSpr2010_c100000059		Marine microbial communities from Delaware Coast, sample from Delaware MO Spring March 2010	79	69861	0.33	91	Bacteria	Proteobacteria
<input type="checkbox"/>	DeMOSpr2010_c10000074	DeMOSpr2010_c100000074		Marine microbial communities from Delaware Coast, sample from	60	64551	0.36	231	Bacteria	Proteobacteria

Example 4:

Now suppose the user wishes to find *Proteobacteria* scaffolds with length greater than or equal to 10,000 bp, and having annotated with both pfam00543 and pfam00909. Since IMG does not allow users to add the same condition filter more than once, the user will have to perform the searches twice: once with pfam00543, and the other time with pfam00909. After each search, the user will have to save the result to a workspace scaffold set, and then get the intersection of both scaffold sets at the end to get the desired result.

However, the user is smart enough to figure out that he can use Pfam ID (for pfam00543) and Pfam Name (for pfam00909) to achieve the result he wants in a single query. That is, the user uses the query builder with the following 4 builder lines:

- Function IDs -- Pfam ID (list) *: pfam00543
- Scaffold Taxonomy -- Phylum: Proteobacteria
- Scaffold Statistics -- Scaffold Nucleotide Length * (Range): >= 10000
- Function Names -- Pfam Name *: Ammonium Transporter

The screenshot displays the IMG query builder interface. At the top, four conditions are added to the query:

- Function IDs -- Pfam ID (list) *: pfam00543
- Scaffold Taxonomy -- Phylum: Proteobacteria
- Scaffold Statistics -- Scaffold Nucleotide Length * (Range): >= 10000
- Function Names -- Pfam Name *: Ammonium Transporter

Below the conditions, the search scope is set to "Selected Genomes". A warning message states: "The total selection for 'Selected Genomes' can only be less than 50 metagenomes. There could be a significant delay if more than 1000 genomes selected or if 'All Isolates' selected." The "Sequencing Status" is set to "All Finished, Permanent Draft and Draft" and the "Domain" is "Genome Cart". The "Selected Genomes" list shows two selected items: 3300000101 and 3300000116. The "Search for" field contains the text "Marine microbial communities from Delaware Coast, sample from Delav". The "Constructed Query" is displayed at the bottom: (Function IDs -- Pfam ID (list) * [pfam00543]) AND (Scaffold Taxonomy -- Phylum [Proteobacteria]) AND (Scaffold Statistics -- Scaffold Nucleotide Length * (Range) [>= 10000]) AND (Function Names -- Pfam Name * [Ammonium Transporter]). The interface includes buttons for "Reset", "Evaluate Query", and "Search". A checkbox for "Hide All Categories" is also present.

After adding the same two metagenome to search and click the **Search** button, the result shows that 19 scaffolds satisfying the search condition:

Advanced Scaffold Search Results

MER-FS Metagenome: assembled

Query:

(Function IDs -- Pfam ID (list) * [pfam00543]): 349 count(s).
 (Scaffold Taxonomy -- Phylum [Proteobacteria]): 360555 count(s). (Earlier results channelled into this search.)
 (Function Names -- Pfam Name * [Ammonium Transporter]): 895 count(s). (Earlier results channelled into this search.)

Selected Genomes: 2

- (Function IDs -- Pfam ID (list) * [pfam00543]): 349 count(s).
- (Scaffold Taxonomy -- Phylum [Proteobacteria]): 360555 count(s). (Earlier results channelled into this search.)
- (Scaffold Statistics -- Scaffold Nucleotide Length * (Range) [>= 10000]): 7375 count(s). (Earlier results channelled into this search.)
- (Function Names -- Pfam Name * [Ammonium Transporter]): 895 count(s). (Earlier results channelled into this search.)

Final Combination: 19 count(s).

[Add Selected to Scaffold Cart](#) [Select All](#) [Clear All](#)

Filter column: [Scaffold ID](#) Filter [text](#) [Apply](#) [?](#)

[Export](#) Page 1 of 1 << first < prev 1 next > last >> [All](#)

[Column Selector](#) [Select Page](#) [Deselect Page](#)

Select	Scaffold ID	Scaffold Name	External Accession	Genome	Gene Count	Sequence Length (bp)	GC Content	Read Depth	Lineage Domain	Lineage Phylum
<input type="checkbox"/>	DeMOSpr2010_c10000002	DeMOSpr2010_c10000002		Marine microbial communities from Delaware Coast, sample from Delaware MO Spring March 2010	137	148253	0.35	207	Bacteria	Proteobacteria
<input type="checkbox"/>	DeMOSpr2010_c10000082	DeMOSpr2010_c10000082		Marine microbial communities from Delaware Coast, sample from Delaware MO Spring March 2010	60	60198	0.38	217	Bacteria	Proteobacteria
<input type="checkbox"/>	DeMOSpr2010_c10000459	DeMOSpr2010_c10000459		Marine microbial communities from Delaware Coast, sample from	25	26406	0.39	66	Bacteria	Proteobacteria

Search History

All the scaffold search history will be recorded in the Search History section in reverse chronicle order:

Search History

- Cart data could be lost when logging out or closing the browser, but can be saved into your [workspace](#).
- For examples, view [public list](#) to save the selected into your cart.

Select	Time	Query	Search In	MER-FS Metagenome		
<input type="checkbox"/>	2020/09/24 14:13:43	(Function IDs -- Pfam ID (list) * [pfam00543]) AND (Scaffold Taxonomy -- Phylum [Proteobacteria]) AND (Scaffold Statistics -- Scaffold Nucleotide Length * (Range) [>= 10000]) AND (Function Names -- Pfam Name * [Ammonium Transporter])	Selected Genomes, 2 genomes	assembled	Reconstruct Query	Search
<input type="checkbox"/>	2020/09/24 14:00:36	(Function IDs -- Pfam ID (list) * [pfam00543, pfam00909]) AND (Scaffold Taxonomy -- Phylum [Proteobacteria]) AND (Scaffold Statistics -- Scaffold Nucleotide Length * (Range) [>= 10000])	Selected Genomes, 2 genomes	assembled	Reconstruct Query	Search
<input type="checkbox"/>	2020/09/24 13:43:42	(Scaffold Statistics -- Scaffold 16S rRNA Gene Count * (Range) [>= 1]) AND (Scaffold Statistics -- Scaffold 23S rRNA Gene Count * (Range) [>= 1])	Selected Genomes, 2 genomes	assembled	Reconstruct Query	Search
<input type="checkbox"/>	2020/09/24 13:29:53	(Scaffold ID (list) [SRS018774_WUGC_scaffold_235])	All Isolates		Reconstruct Query	Search

[Save Selected to Workspace](#)
[Select All](#)
[Clear All](#)
[Remove Selected](#)

Save to Workspace

Search history is like analysis carts: Any queries shown in the history could be lost after users close the browser, and users can save the data to Workspace.

To save any queries to workspace, simply select the queries and click the Save Selected to Workspace button. To view the saved query, go to the **Scaffold Search History** submenu under the Workspace menu item.

Reconstruct Query

The Reconstruct Query button next to each query allows users to view and to revise a previously constructed query.

Rerun Query

The Search button next to the Reconstruct Query button allows users to rerun a previously constructed query.